Respondents' ratings of expressions from response scales: a two-country, two-language investigation on equivalence and translation

Mohler, Peter Ph.; Smith, Tom W.; Harkness, Janet

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Respondents' Ratings of Expressions from Response Scales: A two-country, two-language investigation on equivalence and translation[1]

PETER PH. MOHLER, TOM W. SMITH AND JANET A. HARKNESS

The paper presents German-American research on expressions from response scales used in cross-national and cross-lingual survey research. Respondents in the United States and Germany were asked to rate expression for the degrees of intensity they were held to express. The scales used were scales of agreement, importance and for/against. The findings of the study raise as many questions as they answer. Translation-based pairings of expressions across English and German work well but not perfectly. Symmetrical response scales often lead to artificial-sounding 'scalespeak' constructions: their effect on scale responses is unknown. Well-matched translation pairings were sometimes differently scored across the populations. Germans and Americans differed in the range of scale points they employed and in the range of vocabulary used to 'explain' expressions. The study is seen as a first step towards understanding cross-national response scale issues.

1. Introduction

Cross-national survey research usually takes translated instruments as their route to 'equivalent instruments' (Acquadro, 1996; Van de Vijver, this volume). A number of authors have discussed issues of equivalence and non-equivalence of translated instruments. Others demonstrate and discuss the fact that translation equivalence is only

[1] This research was supported by a grant from the Humboldt Stiftung Transcoop Programme.

one of the equivalencies to be considered in questionnaires (Van de Vijver, this volume; Hui and Triandis, 1986; Hulin, 1987).

The MINTS project (Research into Methodology of Intercultural Surveys) investigated expressions used in response scales in cross-cultural research. The project is the first step in a research programme aimed at exploring the limits and potential of translation with respect to response scales. One of the questions of interest was whether even 'good' translations of expressions used in response scales means that the expressions matched in translation across languages do indeed capture the 'same' or comparable degrees of differentiation. Another was which, if any, of translations already in use for an English response scale would match up best. A further aim was to compare the ratings respondents assigned terms actually used in scales (see section 5.2) with ratings they assigned to terms not used, but potentially usable in scales. And finally, the project investigated what respondents understood various expressions to mean. This is relevant of itself and, we hope, can be linked to corpora research on the various expressions (lexemes) involved.

The MINTS project investigates expressions frequently used in cross-national survey response scales, specifically, expressions used in English and German ISSP[2] response scales. The most commonly used ISSP scales were taken: *agreement/disagreement*; *for/against*; *important/unimportant* (Davis, 1993). Other expressions which are not used in ISSP response scales but are comparable in the degrees of importance, agreement, etc., they express were also investigated (Smith, 1997).

Survey questions generally consist of a (fairly restricted) number of parts which can include an introduction or pre-code, a question-asking part and, in closed format questions, a response scale and instructions such as *Please tick one box*, etc. In the monocultural context, considerable research (not reviewed here) has appeared on almost

[2] The International Social Survey Programme (ISSP) has conducted annual surveys since 1995. Twenty-nine countries are currently members of the ISSP. The data from the survey are distributed by the Zentralarchiv für empirische Sozialforschung in Cologne, Germany.

every aspect of questionnaire design, for example, on *item wording* (e.g., Hippler et al., 1987; Bradburn and Sudmann, 1991; Converse and Presser, 1994; Sudman et al., 1996; Schwarz, 1996), *introductions to questions* (e.g., Cannell et al., 1979; Schumann and Presser, 1981; Converse and Presser, 1994), *length of questions* (e.g., Payne, 1951; Cannell et al., 1979; Converse and Presser, 1994), *question ordering* (e.g., Schumann and Presser, 1981; Hippler et al., 1987; Converse and Presser, 1994; Wänke and Schwarz, 1997; Sudman et al., 1996), and *response scale designs* (e.g., Schumann and Presser, 1981; Presser and Schumann, 1980); Converse and Presser, 1994; Schwarz, 1996; Krosnick and Fabrigar, 1997) to the *interaction between response scales and items* (e.g., Hippler et al., 1987; Schwarz et al., 1991; Schwarz and Hippler, 1991; Schwarz, 1996).

While questions cover a vast range of topics, and there are numerous, albeit 'standard' formats for constructing question-asking parts, response scales, once chosen, tend to be used time and again in identical format. Davis's (1993) review of circa 300 ISSP questions shows the ISSP agreement scale was used 92 times in modules from 1985-1993, the ISSP importance scale, 23 times, an *allow/not allow* scale, 22 times and an *in favour/against* scale, 11 times. Other research programmes, such as the American GSS[3], also repeatedly use the same scales from year to year. Response scales therefore seemed a most useful starting point for our research programme.

2. Agreement Scales across Institutes and Countries

Many major surveys use *agreement* scales and, as just mentioned, often consistently use one or the other format. Where variation occurs, this is often due to taking over questions from other surveys. Across programmes, however, both within one country and across countries, differences in the formulation of a particular scale are frequent. In different

[3] The American General Social Survey (GSS) is an annual survey conducted by the National Opinion Research Center (NORC) in Chicago. The first survey took place in 1972. Further information at the web site (www.norc.uchicago.edu/gss.htm).

programmes in English, for example, one finds the following variations of an agreement scale:

1) A 'forced choice' response scale with only the first two options read out to respondents.

Agree
Disagree
Don't know
No answer
Not applicable

Source: American General Social Survey (GSS), Cumulated Codebook, Q.357a, 1972-1993.

2) A 'forced choice' design using a four-point scale, with two 'agreement' points, two 'disagreement' points and no middle option:

Strongly agree
Agree a little
Disagree a little
Strongly disagree
DK
NA

Source: British Social Attitudes (BSA), Cumulated Sourcebook. K-15 (1987/1989).

3) Seven- or five-point scales provide mid-points and some differentiation in degrees of agreement and disagreement. In addition, the following British scale has the reverse order of modifier to that of the previous scale (italics added here).

Agree *strongly*
Agree
Neither agree nor disagree
Disagree
Disagree *strongly*
DK
NA

Source: British Social Attitudes (BSA), Cumulated Sourcebook. K-15 (1987/1989).

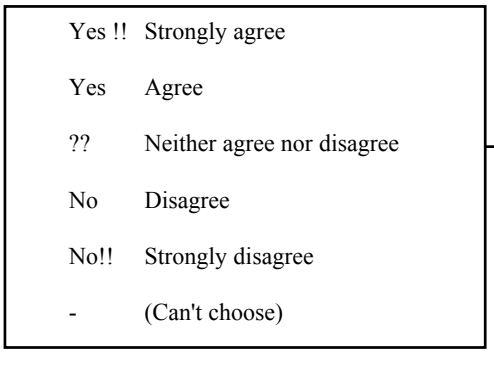
4) The 'standard' ISSP format is as follows (italics added here):

Strongly agree
Agree
Neither agree nor disagree
Disagree
Strongly disagree
Can't choose, Don't know
NA, Refused

Source: ISSP 1993 - GSS (USA) Q 542 A.

5) An Australian version of the standard ISSP scale in example 4) which is used in mail surveys presents the agreement scale and then re-formulates it in terms of *Yes* and *No* and exclamation and question marks, while *Can't choose* seems to become a dash:

To begin with we have some questions about (topic). Do you agree or disagree...(topic)



			Please circle a word			
a. text first item	Yes!!	Yes	??	No	No!!	-
b. text second item	Yes!!	Yes	??	No	No!!	-
c. text third item	Yes!!	Yes	??	No	No!!	-
d. text fourth item	Yes!!	Yes	??	No	No!!	-

Source ISSP 1988 Australia Q 1.

Here both the pre-code wording (*Do you agree or disagree...*) and the scale offered respondents alongside the items differ importantly from the standard ISSP scale.

Cognitive survey methodology research findings show that any one of these differences can affect how respondents react to a scale and the question(s) accompanying it. Numerous findings have demonstrated, for example, that respondents use response scales to interpret questions and questions to interpret scales; that distributions of responses to the 'same' question differ depending on characteristics of the response scales offered; and that the presence or absence of verbal labels or numeric labels, as well as the individual choice of labels, also affect respondents' selection of response options (see Schwarz, 1996, for a review and further references).

Issues of equivalence and the effects of different response scales and scale designs multiply in the cross-national context, in particular when response scales require to be translated. Moreover, the 'close' translation approach often adopted in survey research (Harkness and Schoua-Glusberg, this volume) quickly meets with obstacles in response scale translation. Research on the issues involved is only beginning (Harkness, 1993; 1997; Van de Vijver and Leung, 1997).

3. Measuring the Intensity of Response Categories

The first goal in our research was to establish the *degree* of acceptance, agreement, importance, etc., respondents ascribed to expressions. To do this we needed to measure the degree of intensity respondents assigned to each.

In the monocultural context several approaches have been used to measure the strength of response categories along an underlying response scale. One approach is to have respondents rate the strength of terms defining each point on the scale. There are three standard variants of this approach.

First, one can rank the terms from weaker to stronger or from less to more, or along any similar continuum (cf. Spector, 1976). This, of course, only indicates their relative

position and not the absolute strength or distance between terms. Second, one can rate each term on a numerical scale, usually with 10 to 21 points; (Wildt and Mazis, 1978; Worcester and Burns, 1975; Myers and Warner, 1968; Cliff, 1959; Jones and Thurstone, 1955; Mittelstädt, 1971). This allows the absolute strength or distance between each term to be known and thus facilitates the creation of equal interval scales. It is also possible to use an alphabetical scale or unlabelled spaces, rungs, or boxes, as in a semantic differential scale (Osgood et al., 1957). The letters or spaces are then transformed into their numerical equivalents. Third, magnitude measurement techniques can be used to place each term on a ratio scale (Lodge et al., 1971; 1982; 1992; Wegener, 1991; Hougland et al., 1992). The magnitude measure technique requires that the investigator (sometimes the respondent) give an arbitrary value to a reference term and has respondents rate other terms as ratios to the base term. Typically, respondents have to scale each term by two modes, say, numbers and length of lines. The resulting scales can be calibrated for each individual as well for the whole group of respondents. This allows more precision than the numerical approach, since the terms are not constrained by the artificial limits of the bounded number scale.

Of these three variants, the second seems most useful. On the one hand, the ranking method fails to provide the numerical precision that is necessary to calibrate terms across languages. On the other hand, the magnitude measurement technique is much more difficult to administer and quite difficult for respondents to do, with about 15% of an average population being unable to produce reliable scaling. In addition, the extra precision that a magnitude measurement procedure can provide over that achievable using a 21−point scale approach seems, in our case, to be marginal and thus not needed.

The direct rating approach has been used to rate words along various dimensions. Of most interest here are those that either rate terms along a general good–bad or positive–negative dimension or which rate the intensity of modifiers (Worcester and Burns, 1975; Wildt and Mazis, 1978). Similarly, other studies have rated probability statements (Lichtenstein and Newman, 1967; Wallsten et al., 1986); frequency terms (Simpson,

1944; Spector, 1976; Schaeffer, 1991; O'Muircheartaigh et al., 1993); and terms used in reports to describe percentages from public opinion (Crespi, 1981).

The studies generally show that:

- the tested population (most often American college students) can perform the required rating tasks;
- ratings and rankings are highly similar across different studies and populations (if other than college students);
- there is a high test–retest reliability;
- several different treatments or variations in rating procedures yield comparable results;
- some qualifiers need to be considered differently, as, for instance, vague frequency terms (Schaeffer, 1991; Bradburn and Sudman, 1979).[4]

A second approach for assessing the intensity of scale terms and response qualifiers is to measure the distributions generated by using different response scales (Smith, 1979; Laumann et al., 1994). One version is an *across respondents* design, where two randomly selected groups of respondents get different response scales. With some modelling around what the two observed distributions suggest concerning the supposed underlying distribution, it is possible, within the limits of this approach, to estimate at what point each term cuts the underlying scale (Clogg, 1982; 1984). The assumptions needed for this kind of modeling, namely an underlying 'true' distribution is actually not in line with the more recent literature on judgements and decisions (Schwarz, 1996) or Facet Theory (Borg, 1996; Borg and Groenen, 1997). An alternative version of this approach uses a *within subjects* design. In this, respondents are asked the same question two or more times with different response scales offered (Orren, 1987).

The advantage of the distribution approaches is that they ask respondents to do what they

[4] Experimental settings show systematic differences and artefacts, these seem to vanish or at least to become much smaller in most cases in general population samples and surveys (Weller, 1996).

are normally required to do in the questionnaire context, that is, to answer substantive questions with a standard and typical set of response scales. However, the disadvantages are clear:

– only a very limited number of response scales can be used;

– the statistics need a relatively high number of respondents for each stimulus;

– the implicit model of an underlying 'true' distribution requires detailed analyses.

Since the direct rating approach (asking respondents to rate terms on a 21–point scale) provides the quantified intensity scores needed in the most straightforward manner, this was adopted as the main technique for the MINTS study. At the same time, using a numerical approach in a cross-cultural experiment assumes that respondents in both cultures will respond to and employ numerical values in comparable fashion. While this may be unproblematic for a USA–Germany comparison, in other parts of the world problems are likely, related, for example, to lucky and unlucky numbers, standard (and internalised) rating scales used in education and other spheres, different degrees of familiarity with assessment tasks using more than single digit numbers, etc. These considerations will need to be controlled for in extending our research further.

4. The Study Setting

Experimental pilot studies were carried out in the United States and Germany in 1995 using the direct rating approach described in section 3 to evaluate the equivalence of response scale expressions. The American pilot study was carried out with a sample of adults living in households. Ten sample points were selected to represent all four Census regions (West, South, Midwest, and Northeast). Interviewers had quotas to fill based on gender, age, and employment status. They proceeded through neighbourhoods in the selected communities until the quotas were completed. In contrast to test populations of

college students commonly used in other studies, the respondents of the American pilot study represented the American adult population, according to the stratification variables used for the quota and with respect to marital status and race as a by-product of the selection procedure. Under-represented are, as in many other surveys, the less educated segment of the society. The study was designed and carried out by the National Opinion Research Center at the University of Chicago. Fielding was done in July and August of 1995 with 117 interviews successfully completed (Smith, 1997).

The German experiment was designed as a stand–alone study. By selecting 60 interviewers from different regions, the sample covered all 15 federal states and two main regional substrata, metropolitan regions (100,000 inhabitants and more) and small towns. Within these regional strata, respondents were selected according to a threefold quota table (gender x two age groups x two education groups). The quota cut the population at about the mid-point. As in the American case, the respondents represent the adult population. The sample was split at random to cover two linguistic variants (see below). The study was designed at the German Centre for Survey Research and Methodology (ZUMA); fieldwork was carried out by Infratest-Burke Sozialforschung, Munich. Fieldwork started on September 7 and ended on September 22, 1995. Each interviewer administered only one of the two split-versions; 221 interviews were successfully completed (split 1: 113; split 2: 108).

4.1 Splits

United States: The two American questionnaires differed in question 4 by using *important/unimportant* in one split, and *important/not important* in the other.

Germany: The two German questionnaires differed in all the questions using *agree/disagree* (Q 2, 3, 6, 7, and 8). In split one, *disagree* was translated as *ablehnen*, a verb covering much of the meaning of *disagree/reject*; in split two, *disagree* was translated as ‘not agree’, that is, with *zustimmen* (‘agree’) and a negative particle, *nicht* (‘not’).

4.2 Pairing of English and German Expressions for the Experiment

Selection and pairing of expressions in German to match the English expressions was made on the basis of a) current usage in German surveys, which is itself either based on translations made at some point in time or based on preferred institute or country style, b) translator judgements of appropriateness, and c) formulations which maintained response scale symmetry (Harkness and Mohler, 1997; Harkness, 1993). The experiment was thus able to investigate expressions based on current practice in survey translation and also to expand on this in two relevant directions. All three bases of pairings should be kept in mind when looking at what in some instances might otherwise be surprising alignments.

4.3 Respondents' Ratings of Expressions on a 21-Point Scale

One of the central tasks in the experiment had respondents rate 28 expressions of agreement (26 in English) on a 0 to 20-point scale. Apart from introductory material which contained survey question and answer formats, respondents worked with the expressions outside the survey question-and-answer context. This was important in order to be able at a later stage in research to distinguish between how respondents react to expressions in the questionnaire setting and how they react to these expressions outside of a response scale. Respondents rated each expression in terms of the degree of *agreement/disagreement, importance/unimportance* or 'support for' (in terms of *for/against)* each was felt to express. Theoretically, respondents might be expected to rate *completely agree* somewhere near to 20 and an expression like *completely disagree* near to 0. Respondents were also given the opportunity to adjust their ratings of the agreement expressions once they had completed this task. This revision step was seen as both psychologically useful and informational. It provided some indication of respondent certainty of assessment, gave respondents a chance to look back over the longest and perhaps most demanding task before moving on, and afforded a break in a long sequence of interviewer-respondent dialogue. Respondents did not use this as an opportunity to change ratings to rankings.

4.4 Respondents' Own Definitions of What *Agreement* Means

After the rating part of the experiment, respondents were asked to indicate what they understood the various terms to mean. In English, they were asked the following for agreement: "Now, I'm going to ask you about some of words we've just been discussing. What does the word *agree* meant? What does it involve?" Similar probes were made for *disagree*, *neither agree nor disagree, important*, and *unimportant*. The German respondents were asked as follows: „Im folgenden geht es um einige der Begriffe die Sie gerade eingeordnet haben. Was bedeutet das Wort *stimme zu?* Was heißt das?"

Table 1a below contains the nouns, verbs, adjectives, etc., used by American respondents in their definitions of the meaning of *agree.* Table 1b contains the words used by German respondents in explaining *zustimmen*. Eighty different words were provided by the sixty-one USA respondents taking part in this task. Interviewer records indicate that a fair number of USA respondents used the word asked for as a description of its meaning (e.g., "agree means to agree"). Thirteen of the words used (16,25%) can be seen as variations of the word asked for *(agreement, agreeing),* twenty-five of the words used (31%) can be seen as paraphrases. Of the words used offered by 218 German respondents, 90% of the words chosen can be seen as paraphrases, 6% (fourteen expressions) as repetitions of the word stem of *zustimmen*.

Table 1: 1A - Words used by German Respondents for *zustimmen,* 'agree'

Word	Frequ.	Categ. Frequ.
Akzeptieren	1	
Akzeptabel	1	
Akzeptiere	1	3
Anerkennen	1	
Befürworte	1	
Befürworten	2	
Befürwortung	1	4
Bejahe	1	
Bejahen	3	
Bejahung	4	8
Dafür	42	42
Einverstanden	47	
Einverständnis	7	
Einverständniserklärung	1	55
Gleiche	12	
Gleichen	2	
Gleicher	9	23
Große	1	
Grund	1	
Grunde	3	
Gut	8	8

Word	Frequ.	Categ. Frequ.
Identisch	1	
Positiv	3	
Positive	4	7
Richtig	7	
Richtige	1	
Richtigkeit	1	9
Selbe	4	
Selben	1	
Selber	1	6
Soll	7	
Volle	3	
Volles	1	11
Zustimme	1	
Zustimmen	2	
Zustimmung	11	14
Zutreffend	1	
Zuverlässig	1	
Übereinstimmen	1	
Überzeugt	9	
Übereinstimmung	6	
Überzeugung	3	

1B - Words Used by US-Respondents for *agree*

Word	Frequ.	Categ. Frequ.
About	2	
Accept	3	
Acceptance	2	6
Accomplish	1	
Accord	1	
Accordance	1	2
Admit	1	
Against	1	
Agree	8	
Agreeable	2	
Agreeing	1	
Agreement	2	13
Alike	1	
Approve	3	
Congenial	1	
Consensus	2	

Word	Frequ.	Categ. Frequ.
Consent	2	4
Disagree	1	
Favor	7	
For	5	
Harmony	2	
Like	2	
Liking	1	
Line	2	
Mutual	1	
Ok	1	
Okay	1	
Same	16	
Similar	1	
Support	2	
True	2	
Valid	1	

The readiness of the German respondents to paraphrase or provide alternative expressions and that of Americans to offer the word probed as an explanation of itself can be a reflection of various culturally determined factors (Johnson et al., 1997).

5. Selected Results from the Rating of Agreement Expressions

5.1 *In the Middle* is in the Middle

In the middle and *in der Mitte* both have a mean about the mid-point of the rating scale used. Respondents in both countries not only located expressions such as *neither/nor* and the corresponding German *weder/noch* close to the middle of the scale range, but also placed the so-called 'off-scale' response option of *can't choose* (and *kann ich nicht sagen*

– 'I cannot say') around this middle area, too. Off-scale options are generally understood in survey research as recording the *absenc*e of opinions. It is also sometimes argued that middle categories are used to record non-opinions. Rather than supporting the suggestion that middle options are in fact off-scale options, our findings suggest that middle options, at least in the experimental context, are precisely that. Moreover, expressions implemented in surveys as off-scale options (e.g. *can't choose* and *kann ich nicht sagen)* are in this context close to the centre of the scale, not off-scale (cf. Smith, 1997:13).

Table 2 shows the respective ratings for this middle group of expressions. D stands for the German questionnaire, USA for the American questionnaire, the letters and numbers (e.g., A13 and c in column one) are the respective expression IDS in the two experiments.

Table 2: *In the Middle* and *in der Mitte*

Item IDs D/USA	German Expressions	Mean D	Mean USA	American Expression
A13/c	Stimme ein bißchen zu	12,46	12,10	Agree a little
A26/m	**In der Mitte**	**10,02**	**10,10**	**In the middle**
A22/z	Unentschieden	10,00	9,60	Undecided
A4/p	Stimme weder zu noch lehne ab	9,77	9,90	Neither agree nor disagree
A9/e	Kann ich nicht sagen	9,42	9,80	Can't choose
A7/u	Lehne teilweise ab	6,77	6,60	Somewhat disagree

5.2 'Equivalent' translations do not always have equivalent ratings

Table 3 below shows that *in the middle* and *agree a little*, as well as the German counterparts, *in der Mitte* and *stimme ein bißchen zu,* are rated closer to one another (mean value difference: USA: 2.00 and D: 2.44) than *in the middle/in der Mitte* and the next closest 'disagreement' expression in each language (*disagree a little* (difference

3.00), *lehne teilweise ab* (difference 3.25). Moreover, the distance between *in der Mitte*, *in the middle* to the disagreement expressions which are ‘equivalent’ in terms of word symmetry (*disagree a little* and *lehne ein bißchen ab*) is greater for German (3.05) than for the USA (3.00). The 'structurally equivalent' translation pairing here is not supported by the respondents' ratings. This is suggestive evidence of the dangers of equating linguistic similarity and/or expression symmetry with measurement properties. It may also be related to scalespeak effects, in as much as *disagree a little* is normal English and *lehne ein bißchen ab* is constructed, artificial German.

Table 3: Mean Values of Agree/Disagree Expressions

Item IDs D/US	German Expressions	Mean D	Mean USA	American Expression
A20/v	Stimme voll and ganz zu	19,87	18,80	Strongly agree
A27/f	Stimme völlig zu	19,55	19,40	Completely agree
A17/h	Stimme bestimmt zu	19,22	19,00	Definitely agree
A16/b	Stimme zu	19,05	16,00	Agree
A12/aa	Stimme sehr zu	17,77	18,50	Very much agree
A28/d	Stimme ziemlich zu	16,33	17,20	Agree a lot
A1/a	Stimme im Grunde zu	14,93	13,80	Basically agree
A25/y	Stimme eher zu	13,99	13,50	Tend to agree
A6/r	Stimme wahrscheinlich zu	13,93	13,60	Probably agree
A18/t	Stimme teilweise zu	13,37	12,90	Somewhat agree
A11/n	Stimme mäßig zu	12,49	13,30	Moderately agree
A13/c	Stimme ein bißchen zu	12,46	12,10	Agree a little
A26/m	In der Mitte	10,02	10,10	In the middle
A22/z	Unentschieden	10,00	9,60	Undecided
A4/p	Stimme weder zu noch lehne ab	9,77	9,90	Neither agree nor disagree
A9/e	Kann ich nicht sagen	9,42	9,80	Can't choose
A7/u	Lehne teilweise ab	6,77	6,60	Somewhat disagree
A10/s	Lehne wahrscheinlich ab	6,66	6,20	Probably disagree
A21/o	Lehne mäßig ab	6,63	6,40	Moderately disagree
A24/k	Lehne ein bißchen ab	6,57	7,10	Disagree a little
A19/y	Lehne eher ab	5,82	6,40	Tend to disagree
A15/l	Lehne ziemlich ab	3,91	3,00	Disagree a lot
A14/q	Stimme nicht zu	3,32	3,50	Not agree
A2/i	Lehne bestimmt ab	2,42	1,00	Definitely disagree
A3/j	Lehne ab	2,41	3,50	Disagree
A23/bb	Lehne sehr ab	1,77	1,40	Very much disagree
A5/w	Lehne stark ab	1,21	1,50	Strongly disagree
A8/g	Lehne völlig ab	0,67	0,80	Completely disagree

5.3 Sample Error Variance of Mean Values

In statistical terms, mean values resulting from samples may vary from sample to sample.

Possible variations around a 'true' mean value in the population from which the sample was drawn can be estimated, however. In Figure 1 below, the mean values from our sample are surrounded by vertical lines indicating the band width of stochastically possible variation (variation due to sampling and measurement error – 95% confidence interval). In other words, if the experiment were repeated many times, the expectation is that 95% of the respective mean values would fall within the band width indicated.

Figure 1: Comparison of Means - *agree a little - somewhat disagree* and German counterparts – 95% Confidence Interval

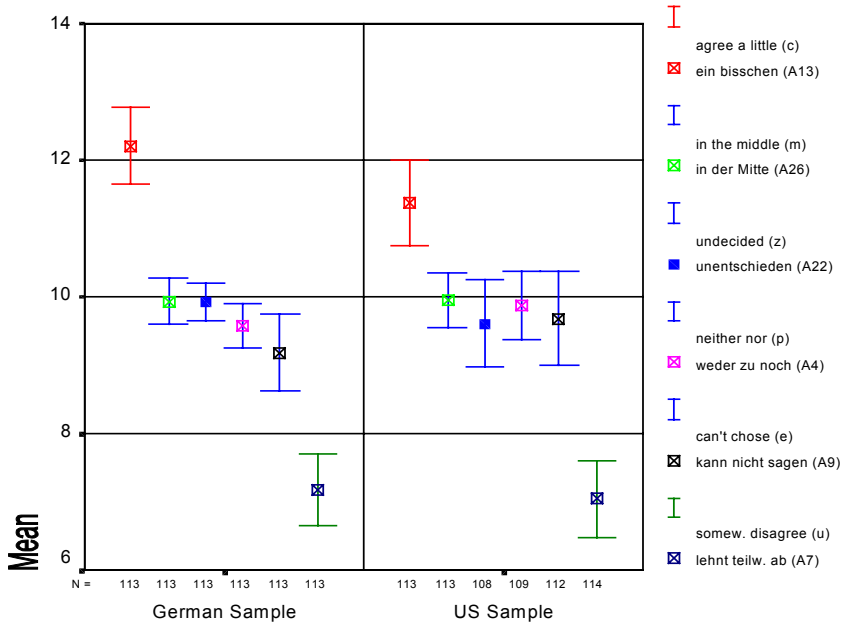


Figure 1 presents these band widths for *agree a little/stimme ein bißchen zu* over *in the middle/in der Mitte* to *somewhat disagree/lehne teilweise ab*. Each vertical bar above and below the boxes indicates the band width of the respective mean value. German results

are plotted on the left, American on the right. A horizontal overlap of the bars indicates that the mean values of the respective expressions are statistically indistinguishable.

The topmost expression here is the agreement pair respondents rated lowest but still above *in the middle/in der Mitte*, that is, *agree a little, stimme ein bißchen zu.* Response pair *lehne teilweise ab/somewhat disagree* is the first *pair* rated below *in the middle*, *undecided*, *neither/nor*, *can't choose* and their German counterparts. We took the German order of mean values here. The first American *expression* in terms of rating is *disagree a little,* as can be seen in Table 3.

The four expressions in each language lexically referring to a mid-point, a non-decision, or an inability to choose, are clustered around the mid-point 10 on the scale. The confidence intervals of the means overlap within countries as well as across, but are distinct from the next 'agreement' and 'disagreement' expressions. In short, the four expressions indicate a mid-point with the same accuracy; they are statistically indistinguishable.

5.4 US and German Differences in Range of Scales

Table 3 findings indicate that the range of scale points American respondents used to rate English expressions is narrower than that used by the German respondents for German expressions. The highest German mean value is 19.87 for *stimme voll und ganz zu*, the American corresponding highest mean value, of 19.40, is for *completely agree*.

On the disagreement ratings, we find a similar pattern. *Lehne völlig ab* is rated as 0.67, while *completely disagree* is located at 0.80. However, inspecting the median values shows this result holds for the top of the scale only (Table 4). This indicates differences across the experiments in dealing with agreement and disagreement which require further investigation.

Table 4: Median Values for *Strongly Agree* and *Strongly Disagree* and *Stimme Voll und Ganz Zu* and *Lehne Völlig Ab*

German expression	Median, Germany	Median, America	American expression
Stimme voll and ganz zu	20,0	19,0	Strongly agree
Stimme zu	18,0	16,5	Agree
Weder zustimmen/noch ablehnen	10,0	10,0	Neither agree nor disagree
Lehne ab	03,0	03,5	Disagree
Lehne stark ab	01,0	01,0	Strongly disagree

6. Summary of Main Findings

The rating experiments showed in general a high correspondence between the *a priori* pairings of expressions by researchers in the United States and Germany. Most means are close and not statistically different from one another (Mohler et al., 1996). Despite this extremely high correspondence, expressed in correlation coefficients above 0.9 (Smith, 1997), there are, nevertheless, some important differences in the mean values. First, the simple base terms such as *agree – stimme zu*, *disagree – lehne ab/stimme nicht zu* are rated more extremely by German respondents than their English counterparts are by American respondents. It remains to be seen whether this means the German expressions involve greater intensity of agreement/disagreement, etc. or whether, independent of this, German respondents differ in rating behaviour. Certainly, in other languages and cultures, response behaviour and the intensity of agreement/disagreement associated with unmodified base terms do seem to differ (Johnson et al., 1997).

Some expressions rank differently across the two countries. Thus in the US experiment, respondents gave the following order to expressions (in the middle = 1):

US Sequence No.1	2	3	4
in the middle	*disagree a little*	*somewhat disagree*	*moderately disagree & tend to disagree*
'German Pair' Sequence No.1	5	2	4

In Germany the expressions paired to the above by researchers were ordered by respondents as follows:

German Sequence No.1	2	3	4
in der Mitte ('in the middle')	*lehne teilweise ab ('disagree/reject in part')*	*lehne wahrscheinlich ab ('probably disagree/ reject')*	*lehne mäßig ab ('moderately disagree')*
'US Pair' Sequence No.1	3	5	4

Differences in ranking in the two populations can be noted for scalespeak pairs such as *disagree a little* and (scalespeak) *lehne ein bißchen ab* but also for expressions which, at face value, are well-paired, ordinary translatory equivalents *(lehne wahrscheinlich ab, probably disagree).*

7. The Next Steps

Assessment of response scales in translation can neither be limited to assessment of translating equivalence (however defined, cf. Harkness and Schoua-Glusberg, this volume; Harkness and Braun, in preparation) nor assessment of measurement properties.

For instance, the effects of scalespeak characteristics across languages have, to our knowledge, never been investigated. If, for example, symmetrical scalespeak designs skew response scales, then other expressions which do not observe scale symmetry might be preferable. Moreover, linguistic corpora could be used to provide researchers with a wider range of expressions to choose from; these could, moreover, be evaluated in their habitual or preferred contexts. In this way, researchers would have concrete evidence of whether, for example, a modifier is usually used with positive or negative headwords or whether headwords are gradable (potentially a part explanation of why *a little bit unimportant* is unusual).

Our findings are based on respondents' reactions to expressions removed from the answer scale context. It remains to be seen to what extent these carry over to a response scale context. On the basis of our findings, for example, the expressions used in the English ISSP agreement scale are not equidistant from one another in the degrees of agreement/disagreement respondents felt they expressed. The same applies to the expressions used in German as standard response scale translations of these. We now plan to test respondents' reactions to standardly used response scales against their reactions to response scales using other expressions which our findings indicate might signal more equidistant intervals.

References

Acquadro, C., Jambon, B., Ellis, D. and Marquis, P. (1996). Language and Translation Issues. In: B. Spilker, (ed.), *Quality of Life and Pharmacoeconomics in Clinical Trials* (2nd edition). Philadelphia: Lippincott-Raven.

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling. Theory and Applications.* New York: Springer.

Bradburn, N.M. and Sudman, S. (1979*). Improving Interview Method and Questionnaire Design.* San Francisco: Jossey-Bass.

Bradburn, N.M. and Sudmann, S. (1991). The Current Status of Questionnaire Design. In: P.P. Biemer, et al. (eds.), *Measurement Errors in Surveys* (pp. 29-40). New York: John Wiley & Sons.

Cannel, Ch.F., Oksenberg, L. and Converse, J.M. (1979). *Field Experiments in Health Reporting 1971-1977.* ISR Research Report Series. Ann Arbor: ISR.

Cliff, N. (1959). Adverbs as Multipliers. *Psychological Review* 66: 27−44.

Clogg, C.C. (1982). Using Association Models in Sociological Research: Some Examples. *American Journal of Sociology*, 88: 114−134.

Clogg, C.C. (1984). Some Statistical Models for Analyzing Why Surveys Disagree. In: Ch.F. Turner and E. Martin (eds.), *Surveying Subjective Phenomena*. Vol. 2. New York: Russel Sage.

Converse, J.M. and Presser, S. (1994). Survey Questions: Handcrafting the Standardized Questionnaire. In: M.S. Lewis-Beck (ed.), *Research Practice* (pp. 89-162). London: Sage/Toppan.

Crespi, L.P. (1981). *Semantic Guidelines to Better Survey Reportage.* Office of Research, International Communication Agency, Memorandum.

Davis, J.A. (1993). Memorandum to the ISSP, Chicago: NORC (mimeo).

Harkness, J.A. (1996a). Mountains and Molehills - Equivalence in Cross-Cultural surveys: the Case of Response Scales. (Based on a paper first presented at the American Association of Public Opinion Research, St Charles, 1993).

Harkness, J.A. (1996b). The Representation of Selves in Everyday Questionnaires. Paper presented at the 4th International Sociological Association Conference on Survey Methodology, Colchester, England.

Harkness, J.A. and Mohler, P.Ph.(1997). Towards a Manual of European Background Variables. ZUMA Report on Background Variables in a Comparative Perspective. ZUMA: Mannheim (mimeo).

Hippler, H.-J., Schwarz, N. and Sudman, S. (1987). *Social Information Processing and Survey Methodology.* Heidelberg: Springer.

Hougland, J.G., Johnson, T.P. and Wolf, J.G. (1992). A Fairly Common Ambiguity: Comparing Rating and Approval Measures of Public Opinion. *Sociological Focus 25*: 257-271.

Hui, C.H. and Triandis, H.C. (1985). Measurement in Cross-Cultural Psychology. A Review and Comparison of Strategies. *Journal of Cross-Cultural Psychology 16(2)*: 131-152.

Hulin, C.L., Drasgow, F. and Komocar, J. (1982). Applications of Item Response Theory to Analysis of Attitude Scale Translations. *Journal of Applied Psychology 67(6)*: 818-825.

Hulin, C.L. (1987). A Psychometric Theory of Evaluations of Item and Scale Translations: Fidelity Across Languages. *Journal of Cross-Cultural Psychology 18(2)*: 115-142.

Jones, L.V. and Thurstone, L.L. (1955). The Psychophysics of Semantics: An Experimental Investigation. *Journal of Applied Psychology 39*: 31-36.

Johnson, T., O'Rourke, D., Chavez, N., Sudman, S., Warnecke, R., Lacey, L. and Horn, J. (1997). Social Cognition and Responses to Survey Questions among Culturally Diverse Populations. In: L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewim (eds.), *Survey Measurement and Process Quality* (pp. 87-113). New York: John Wiley & Sons.

Krosnick, J.A. and Fabrigar, L.A. (1997). Designing Rating Scales for Effective Measurement in Surveys. In: L. Lyberg, et al. (eds.), *Survey Measurement and Process Quality* (pp. 141-164). New York: John Wiley & Sons.

Laumann, E.O., Gagnon, J.H., Michael, R.T. and Michaels, S. (1994). *The Social Organization of Sexuality: Sexual Practices in the United States.* Chicago: University of Chicago Press.

Lichtenstein, S. and Newman, J.R. (1967). Empirical Scaling of Common Verbal Phrases Associated with Numerical Probabilities, *Psychon. Sci. 9*: 563-564.

Mohler, P.Ph., Harkness, J.A., Smith, T.W. and Davis, J.A. (1996). Calibrating Response Scales Across Two Languages and Cultures. ZUMA: Mannheim (mimeo).

Mittelstaedt, R.A. (1971). Semantic Properties of Selected Evaluative Adjectives: Other Evidence. *Journal of Marketing Research 8*: 236-237.

Myers, J.H. and Warner, W.G. (1968). Semantic Properties of Selected Evaluation Adjectives. *Journal of Marketing Research 5*: 409-412.

Payne, S.L. (1951). *The Art of Asking Questions.* Princeton/NJ: Princeton University Press.

O'Muircheartaigh, C.A., Gaskell, G.D. and Wright, D.B. (1993). The Impact of Intensifiers. *Public Opinion Quarterly 57*: 552-565.

Orren, G.R. (1978). Presidential Popularity Ratings: Another View. *Public Opinion 1*: 35.

Osgood, Ch.E., Suci, G.J. and Tannenbaum, P.H. (1957). *The Measurement of Meaning.* Urbana, IL: University of Illinois Press.

Presser, S. and Schumann, H. (1980)

Schaeffer, N.C. (1991). Hardly Ever or Constantly? Group Comparisons Using Vague Quantifiers. *Public Opinion Quarterly 55*: 395-423.

Schönemann, P.H. (1994). Measurement: The Reasonable Ineffectiveness of Mathematics in the Social Sciences. In: I. Borg and P.Ph. Mohler (eds.), *Trends and Perspectives in Empirical Social Research*. (pp. 149-160). Berlin: de Gruyter.

Schumann, H. and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context.* New York: Academic Press.

Schwarz, N. (1996). *Cognition and Communication.* Mahawah/NJ: Lawrence Erlbaum.

Schwarz, N. and Hippler, H.-J. (1991). Response Alternatives: The Impact of their Choice and Presentation Order. In: P.P. Biemer et al. (eds.), Measurement Errors in Surveys. (pp. 41-56). New York: John Wiley & Sons.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E. and Clark, L. (1991). Rating Scales: Numeric values may change the meanings of scale labels. *Public Opinion Quarterly 55*: 570-582.

Simpson, R.H. (1944). The Specific Meanings of Certain Terms Indicating Differing Degrees of Frequency. *Quarterly Journal of Speech 30*: 328-330.

Smith, T.W. (1979). Happiness: Time trends, seasonal variations, intersurvey differences, and other mysteries. *Social Psychology Quarterly 42*: 18-30.

Smith, T.W. (1997). Improving Cross-National Survey Research by Measuring the Intensity of Response Categories. GSS Cross-National Report No. 17. Chicago: NORC (mimeo).

Spector, Paul E., (1976). Choosing Response Categories for Summated Rating Scales. *Journal of Applied Psychology 61:* 374−375.

Stone, L. and Campbell, J. (1984). The Use and Misuse of Surveys in International Development: An Experiment from Nepal. *Human Organization 43*: 30-37.

Sudman, S., Bradburn, N.M. and Schwarz, N. (1996). Thinking About Answers – *The Application of Cognitive Processes to Survey Methodology.* San Francisco: Jossey-Bass.

Van de Vijver, F.J.R. and Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research.* Newbury Park/CA: Sage.

Wänke, M. and Schwarz, N. (1997). Reducing Question Order Effects: The Operation of Buffer Items. In: L. Lyberg et al. (eds.), *Survey Measurement and Process Quality* (pp. 115-140). New York: John Wiley & Sons.

Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R. and Forsyth, B. (1986). Measuring the Vague Meanings of Probability Terms. *Journal of Experimental Psychology 115*, 348–365.

Wegener, B. (ed.) (1991). *Social Attitudes and Psychophysical Measurement.* Hillsdale/NJ: Lawrence Erlbaum.

Weller, I. (1996). Kontexteffekte in Eurobarometer Umfragen - Theoretische Implikationen und praktische Bedeutung. Unpublished dissertation University of Heidelberg.

Wildt, A.R. and Mazis, M.B. (1978). Determinants of Scale Response: Label vs. Position. *Journal of Marketing Research 15*: 261–267.

Worcester, R.M. and Burns, T.R. (1975). A Statistical Examination of the Relative Precision of Verbal Scales. *Journal of the Market Research Society 17*: 181–197.